Let X, Y, and K be the stochastic variables associated with x, y, and k. Let $H(y)$ be the entropy of Y when nothing is known about k. Let $H_x^{(i)}(y)$ be the conditional entropy of Y for a given x, after i pairs (x_j,y_j) have been intercepted. Let $H(k)$ be the entropy of K. Since $f(x,k)$ is an unknown function if k is unknown, this equation from [2] is valid:

$$\sum_{i=0}^{\infty} H_x^{(i)}(y) \leqslant H(k). \tag{1}$$

The number of keys is finite, although large, so the sum in (1) is also finite. Then most terms are very small or zero. Let p_i be the maximum probability that C chooses a key k that yields the correct y' for x' when i pairs (x,y) have been intercepted. A slight modification of the lemma in [1, p. 410] yields

$$p_i \geqslant 2_x^{-H(i)}(y) \tag{2}$$

with equality if and only if $y|x$ has a rectangular density function which is independent of the value of x.

A very small $H_x^{(i)}(y)$ then means that p_i is close to one. In practice, we are hardly interested in having an authentication function f that makes it highly likely that C will succeed in substituting his x' for most of the x_i. So f is most sensibly constructed in such a way that $H_x^{(i)}(y)=0$, if $i \geqslant N$. Then the rest of the $H_x^{(i)}(y)$ can be made as large as possible, thereby making it possible to have p_i small for $i<N$. Such an f will be the best choice for exactly N messages. If there are more than N messages then $p_i=1$ for the last ones, which amounts to total risk of fraud. If only $n<N$ messages are sent, a slightly better bound for p_i in (2) could have been obtained with another f.

We now want to maximize the security of N messages, i.e., N samples of the stochastic variable X, which means that we want to minimize the average risk

$$P=\frac{1}{N}\sum_{i=0}^{N-1} p_i. \tag{3}$$

Equations (2) and (3) yield

$$P \geqslant \frac{1}{N}\sum_{i=0}^{N-1} 2_x^{-H(i)}(y) \tag{4}$$

with equality if and only if the variables $y|x$ for $i=0,1,\cdots,N-1$ have a rectangular density function. Thus f should be constructed to yield such a density function.

For simplicity, all intercepted pairs are supposed to be different, i.e., no x is sent twice. Also $p_N=1$, since $H_x^{(N)}(y)=0$. Let L_i denote the number of keys that are still possible choices for the correct key after i interceptions of pairs $(x,y=f(x,k))$. Thus $L_0=L$ and $L_N=1$. (If two keys k_i and k_j have $f(x,k_i)=f(x,k_j)$ for all x, then k_i and k_j are considered as the same key. L is the number of different functions f that the set of keys can produce.)

We now suppose that the keys have a rectangular distribution function, since any other distribution would be more advantageous to C. Furthermore, we suppose that f is constructed to yield equality in (4). Then L_i is independent of both the chosen key and the sequence of received x_j, and so is L_{i+1} independently of the value of x_{i+1}. So of all L_i keys that remain after i interceptions L_{i+1} can not be ruled out after the next interception, since they all produce the same correct y_{i+1} for the given x_{i+1}. Thus L_{i+1} keys can be safely used to produce a correct y_i'. But C doesn't know which one to choose from the L_i possibilities. This means that

$$p_i=\frac{L_{i+1}}{L_i}. \tag{5}$$

So $p_0=L_1/L$ and $p_{N-1}=1/N$. Since $L_{i+1}=p_i \cdot L_i$, we also get

$$L_N=1=L\prod_{i=0}^{N-1} p_i. \tag{6}$$

Thus we want to minimize P under the condition (6), which can be rewritten as

$$\prod_{i=0}^{N-1} p_i=\frac{1}{L}. \tag{7}$$

We now use Lagrange multipliers to choose the p_i so as to minimize P in (3) subject to the constraint (7). The answer is that we should choose

$$p_i=L^{-1/N} \tag{8}$$

and then

$$\min(P)=\min\left(\frac{1}{N}\sum_{i=0}^{N-1} p_i\right)=L^{-1/N}. \tag{9}$$

This is hardly a surprising result. To check it, consider the cases $N=1$ and $2.N=1$ is the situation when x_1 is a password which will never be intercepted, but C nevertheless tries to impersonate A. Equations (8) and (9) state correctly that the probability of success for C is $1/L$ if all the passwords are equally probable, and this probability is a minimum. $N=2$ is the case treated in [1]. Equations (8) and (9) state that the probability of success for C is $L^{-1/2}$, which is the result obtained in [1].

The consequences of (8) and (9) for f are that if we retain the picture introduced in [1, Fig. 2] of bundles of keys leading from each x to different y, then we get the following rule. If we choose i different x and one key at random, consider the i bundles of keys which lead from these x and contain the chosen key. There should then be exactly $L^{(N-i)/N}$ keys, including the chosen one, which appear in all i bundles. Another way to put this is to say that the intersection of i different $G(x_j,y_j)$ should contain $L^{(N-i)/N}$ keys.

References

[1] E. N. Gilbert, F. J. MacWilliams, and N. J. A. Sloane, "Codes which detect deception," *Bell Syst. Tech. J.*, vol. 53, no. 3, pp. 405–424, Mar. 1974.

[2] I. Ingemarsson, "Toward a theory of unknown functions," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 238–240, March 1978.

Convolutional Encoding for Wyner's Wiretap Channel

ERIK VERRIEST AND MARTIN E. HELLMAN,
SENIOR MEMBER, IEEE

***Abstract*—The wiretap channel introduced by Wyner is studied for the special case when the main channel is a noiseless binary channel and the wiretap channel is a binary symmetric channel. With a rate-one convolutional encoder, the steady-state uncertainty of the wiretapper is shown to depend only on the constraint length ν of the code, not on the specific taps, and is complete on k successive bits provided $k \leqslant \nu$. During the initial transient period, the rate of growth of uncertainty does depend on the tap connections of the shift register.**

I. Introduction

The wiretap channel, introduced by Wyner [1], is an interesting case of a broadcast channel [2], [3] in which the information flow to one receiver is to be maximized while the information

Manuscript received September 26, 1977; revised May 10, 1978. This work was supported in part by the U.S. Air Force Office of Scientific Research under Contract F44620-73-C-0065 and in part by the Joint Services Electronics Program under Contract N00014-75-C-0601.

The authors are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305.

0018-9448/79/0300-0234$00.75

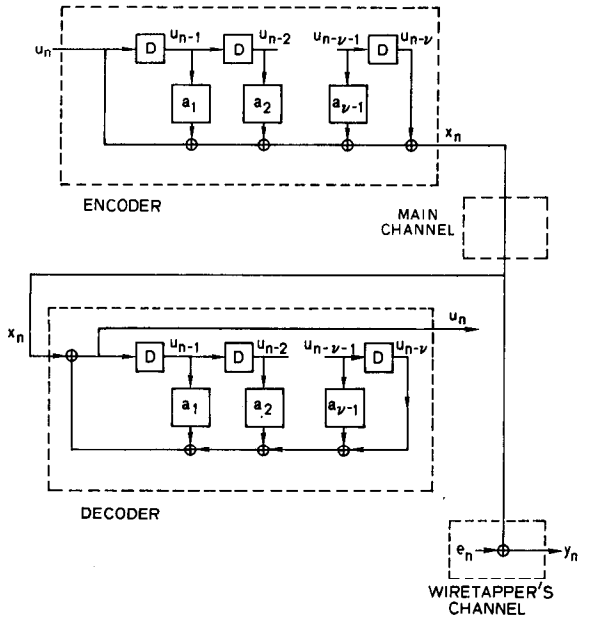

Fig. 1. Convolutional encoder, feedback decoder, and wiretapper.

flow to the other receiver is to be minimized. The first receiver is referred to as the legitimate receiver, and the second receiver is referred to as the wiretapper. The wiretap channel and variations thereof have received considerable attention [4]–[9] with the use of block codes. This note examines the use of convolutional codes for the special case where the main channel to the legitimate receiver is a noiseless binary channel and the wiretapper's channel is a binary symmetric channel (BSC) with bit error rate ϵ, as depicted in Fig. 1.

Because the main channel is noiseless, the encoder can operate at rate one and is only used to compound the wiretapper's uncertainty. This additional uncertainty does not result from lack of knowledge of the code, which is assumed to be publicly known. Rather, it results from memory in the decoder, which allows a small amount of uncertainty on each of the preceeding bits to concentrate in one portion of the message. The encoder and the decoder used at the legitimate receiver are shown in Fig.1. Note that, without loss of generality, a_0 and a_ν are assumed equal to one.

The initial state of the encoder, $s_e(0)=(u_0,u_{-1},\cdots,u_{-\nu+1})$, is also assumed to be public information, and the decoder is started in the same initial state $s_d(0)=s_e(0)$. By linearity, we may take this common initial state to be $\mathbf{0}$ without loss of generality.

We are interested in evaluating the wiretapper's uncertainty concerning portions of the information sequence $\boldsymbol{u}$ after he has observed all of y. Letting

$$\boldsymbol{u}_{n+1}^{n+k}=(u_{n+1},u_{n+2},\cdots,u_{n+k}), \tag{1}$$

$$\boldsymbol{u}^n=\boldsymbol{u}_1^n, \tag{2}$$

and $\boldsymbol{u}=\boldsymbol{u}^\infty$ and using similar notation for other sequences, we wish to evaluate

$$H(\boldsymbol{U}_{n+1}^{n+k}|\boldsymbol{Y}) \tag{3}$$

under the assumption that each u_i is an independent Bernoulli (1/2) random variable.

Considering the case $n=0$, it is seen from Fig. 1 that there is a 1:1 correspondence between $\boldsymbol{x}^k$ and $\boldsymbol{u}^k$ and that $\boldsymbol{y}_{k+1}^\infty$ is independent of $\boldsymbol{x}^k$. Therefore

$$H(\boldsymbol{U}^k|\boldsymbol{Y})=H(\boldsymbol{X}^k|\boldsymbol{Y}^k) \tag{4}$$

$$=H(\boldsymbol{E}^k) \tag{5}$$

$$=kh(\epsilon) \tag{6}$$

where $\boldsymbol{E}$ is the random error sequence on the wiretap channel and

$$h(\epsilon)=-\epsilon\log_2\epsilon-(1-\epsilon)\log_2(1-\epsilon) \tag{7}$$

is the binary entropy function. Therefore, when $n=0$ in (3), the wiretapper's uncertainty is the same as if no encoding were performed ($x_n=u_n$). We shall see, however, that, for large values of n (when the effect of the known initial state is unimportant) and for $k\leqslant\nu$, encoding is extremely useful in increasing the wiretapper's uncertainty about $\boldsymbol{u}_{n+1}^{n+k}$.

II. Steady-State Behavior

Lemma 1: Let $\boldsymbol{d}$ denote the response of the decoder to the sequence of errors $\boldsymbol{e}$. Then

$$H(\boldsymbol{U}_{n+1}^{n+k}|\boldsymbol{Y})=H(\boldsymbol{D}_{n+1}^{n+k}). \tag{8}$$

Proof: There is a 1:1 correspondence between $\boldsymbol{y}$ and $\hat{\boldsymbol{u}}(\boldsymbol{y})$, the response of the decoder to the $\boldsymbol{y}$ sequence, so that

$$H(\boldsymbol{U}_{n+1}^{n+k}|\boldsymbol{Y})=H(\boldsymbol{U}_{n+1}^{n+k}|\hat{\boldsymbol{U}}) \tag{9}$$

and, by linearity,

$$\hat{\boldsymbol{u}}=\boldsymbol{u}\oplus\boldsymbol{d},$$

and hence

$$H(\boldsymbol{U}_{n+1}^{n+k}|\hat{\boldsymbol{U}})=H(\boldsymbol{D}_{n+1}^{n+k}). \tag{10}$$

Lemma 2:

$$\lim_{n\to\infty} H(\boldsymbol{U}_{n+1}^{n+\nu}|\boldsymbol{Y})=\nu \tag{11}$$

Proof: By Lemma 1, we are concerned with

$$H(\boldsymbol{D}_{n+1}^{n+\nu})=H(\boldsymbol{S}_d(n+\nu)) \tag{12}$$

where $\boldsymbol{s}_d(i)=(d_i,d_{i-1},\cdots,d_{i-\nu+1})$ is the state of the decoder at time i when $\boldsymbol{u}=\mathbf{0}$ (i.e., when driven by the $\boldsymbol{e}$ sequence). The state sequence is a first-order Markov chain with 2^ν states and state transition matrix

$$P=\epsilon P^1+(1-\epsilon)P^0 \tag{13}$$

where P^0 and P^1 are the deterministic state transition matrices under $e=0$ and $e=1$, respectively. Because each state has one predecessor under $e=0$ and one predecessor under $e=1$, P is a doubly stochastic matrix (each column as well as each row sums to 1). The limiting state distribution vector $\boldsymbol{\mu}$ must be a solution [10, p. 248] to

$$\boldsymbol{\mu}=\boldsymbol{\mu}P, \tag{14}$$

and because P is doubly stochastic, one solution is

$$\boldsymbol{\mu}=(1/2^\nu)\mathbf{1}. \tag{15}$$

Further, this is the only solution because each state is reachable from every other state [10, p. 251]. Hence the 2^ν possible states are equidistributed in the limit as $n\to\infty$ and the state entropy tends to ν.

Theorem 1:

$$\lim_{n\to\infty} H(\boldsymbol{U}_{n+1}^{n+k}|\boldsymbol{Y})=\begin{cases}k, & \text{if } k\leqslant\nu,\\ \nu+(k-\nu)h(\epsilon), & \text{if } k>\nu.\end{cases} \tag{16}$$

Proof: Lemma 2 establishes the theorem for $k=\nu$ and hence for $k\leqslant\nu$. (If a sequence of k bits has maximal uncertainty, so must any subset.) From Lemma 1

$$H(\boldsymbol{U}_{n+1}^{n+k}|\boldsymbol{Y})=H(\boldsymbol{D}_{n+1}^{n+k}). \tag{17}$$

When $k>\nu$,

$$H(\boldsymbol{D}_{n+1}^{n+k})=H(\boldsymbol{D}_{n+1}^{n+\nu})+H(\boldsymbol{D}_{n+\nu+1}^{n+k}|\boldsymbol{D}_{n+1}^{n+\nu}) \tag{18}$$

$$=H(\boldsymbol{D}_{n+1}^{n+\nu})+H(\boldsymbol{D}_{n+\nu+1}^{n+k}|\boldsymbol{S}_d(n+\nu))$$

$$=H(\boldsymbol{D}_{n+1}^{n+\nu})+(k-\nu)h(\epsilon). \tag{19}$$

Taking the limit as $n\to\infty$ and applying Lemma 2 yields the desired result:

$$\lim_{n\to\infty} H(\boldsymbol{U}_{n+1}^{n+k}|\boldsymbol{Y}) = \nu + (k-\nu)h(\epsilon). \tag{20}$$

It is somewhat surprising that the steady-state behavior is independent of the tap connections, $a_1, a_2, \cdots, a_{\nu-1}$. It is not surprising, though, that the wiretapper's fractional uncertainty is greatest on blocks of length ν or less and drops toward $h(\epsilon)$ on blocks much longer than ν.

III. Transient Behavior

The wiretapper's uncertainty about $\boldsymbol{u}_{n-\nu+1}^{n}$ is equal to $H(\boldsymbol{S}(n))$ when the decoder is driven from $\boldsymbol{s}(0)=\boldsymbol{0}^{\nu}$ by an independent Bernoulli (ϵ) sequence. While the steady-state behavior is independent of the tap connections, this is not true of the transient behavior.

Unfortunately, the transient behavior cannot be simply categorized. The maximal length shift register (MLSR) taps demonstrate a rapid growth of $H(\boldsymbol{S}(n))$, but numerical evaluation showed that they do not always maximize $H(\boldsymbol{S}(n))$ for every n.

It is possible to show, however, that, as $\epsilon\to 0$ or as $\epsilon\to 1$, the MLSR taps maximize $H(\boldsymbol{S}(n))$ for any fixed value of n. As $\epsilon\to 0$, the entropy $H(\boldsymbol{S}(n))$ is dominated by the occurrence of a single one in the $\boldsymbol{e}^n$ sequence. After a single one followed by all zeros, the feedback shift register goes into a cycle of length l_1. For an MLSR, $l_1 = 2^{\nu}-1$, while for any other tap connections, $l_1 < 2^{\nu}-1$. The uncertainty of the phase of $\boldsymbol{S}(n)$ given that a single one (error) has occurred is therefore greatest for an MLSR.

The slowest rate of growth of $H(\boldsymbol{S}(n))$ has been found to occur for $a_1 = a_2 = \cdots a_{\nu-1} = 0$ for $\nu \leqslant 6$, and we conjecture that this is always the case. We have established the following theorems, but in general the transient behavior appears difficult to characterize.

Theorem 2: $H(\boldsymbol{S}(n))$ is invariant to $\boldsymbol{s}(0)$ and under ϵ being changed to $1-\epsilon$.

Proof: By linearity, the state sequence is the sum (modulo-two) of the zero-input response to the initial state and the zero-state response to the input. The zero-input response to the initial state is deterministic and does not affect $H(\boldsymbol{S}(n))$.

Similarly, if $\boldsymbol{e}'$ is Bernoulli $(1-\epsilon)$, then

$$\boldsymbol{e} = \boldsymbol{1} \oplus \boldsymbol{e}' \tag{21}$$

is Bernoulli (ϵ). By linearity, the zero-state response to $\boldsymbol{e}$ is therefore the sum of the responses to $\boldsymbol{1}$ and $\boldsymbol{e}'$. Because the response to input $\boldsymbol{1}$ is deterministic, $H(\boldsymbol{S}(n))$ is the same for inputs $\boldsymbol{e}$ and $\boldsymbol{e}'$, with parameters ϵ and $1-\epsilon$.

Theorem 3: When $a_1 = a_2 = \cdots = a_{\nu-1} = 0$,

$$H(\boldsymbol{S}(k\nu+j)) = (\nu-j)h(\epsilon^{(*k)}) + jh(\epsilon^{(*(k+1))}) \tag{22}$$

for $k \geqslant 0$ and $0 \leqslant j \leqslant \nu$, where

$$\epsilon^{(*k)} = \epsilon^{(*(k-1))} * \epsilon \tag{23}$$

$$\epsilon_1 * \epsilon_2 = \epsilon_1(1-\epsilon_2) + (1-\epsilon_1)\epsilon_2 \tag{24}$$

and

$$\epsilon^{(*1)} = \epsilon. \tag{25}$$

Proof: Equation (24) says that $\epsilon_1 * \epsilon_2$ is the bias of a Bernoulli random variable which is the modulo-two sum of two independent Bernoulli random variables with biases ϵ_1 and ϵ_2. Equation (23) then says that $\epsilon^{(*k)}$ is the bias of the sum of k independent Bernoulli (ϵ) random variables.

When $a_1 = a_2 = \cdots = a_{\nu-1} = 0$ and $0 < j \leqslant \nu$, Fig. 1 shows that

$$d_{k\nu+j} = e_{k\nu+j} \oplus e_{(k-1)\nu+j} \oplus \cdots \oplus e_j \tag{26}$$

so $d_{k\nu+j}$ has a Bernoulli $(\epsilon^{(*(k+1))})$ distribution. Similarly, the first j of $d_{k\nu+j}, d_{k\nu+j-1}, \cdots, d_{(k-1)\nu+j+1}$ are Bernoulli $(\epsilon^{(*(k+1))})$ and the last $(\nu-j)$ are Bernoulli $(\epsilon^{(*k)})$ random variables. Because no e_n enters into more than one sum of the form (26), these random variables are independent. Thus

$$H(\boldsymbol{S}(k\nu+j)) = H(\boldsymbol{D}_{(k-1)\nu+j+1}^{k\nu+j}) = (\nu-j)h(\epsilon^{(*k)}) + jh(\epsilon^{(*(k+1))}) \tag{27}$$

as claimed.

The transient behavior is not too important for long messages where a sequence of random bits can precede the actual information, much as a trailer is used in error correcting convolutional codes. Better characterization of the transient behavior would be of interest, however, because it is a fundamental property of shift registers and likely to find other applications.

IV. Discussion

If a rate-one convolutional encoder of the type studied is used to confuse a wiretapper and if his bit error rate ϵ is small, the constraint length ν must be large enough to prevent the wiretapper from searching over the 2^{ν} typical sequences left after he has received $\boldsymbol{y}$. This dictates that ν be at least 100.

It is interesting to contrast the behavior of the convolutional code studied with that of the feedback code obtained by interchanging the encoder and decoder in Fig. 1. The wiretapper's uncertainty still is equal to the entropy of the (new) decoder's response $\boldsymbol{d}$ to the error sequence $\boldsymbol{e}$. But now d_n depends only on $\boldsymbol{e}_{n-\nu}^{n}$, and if $(\nu+1)h(\epsilon) < 1$, the wiretapper's uncertainty on even single bits will not tend to one as $n\to\infty$. While this system is of less direct value, the properties of the $\boldsymbol{d}$ sequence are very interesting. As indicated by Shepp and Slepian [11], it is not Markov of any finite order, yet is "almost Markov" in many ways. For example, $I(D_n; D_{n+k}|\boldsymbol{D}_{n+1}^{n+k-1})$ tends to zero as $k\to\infty$. In steady state, it is a B-process [12].

A final word of caution is in order. If the information sequence $\boldsymbol{u}$ is not totally random, but possesses redundancy, then either encoding operation may lower the wiretapper's uncertainty, rather than raise it. This is because the encoding operation then serves as a convolutional joint source-channel encoder [13] and, if $H(U) < 1 - h(\epsilon)$, the wiretapper can recover $\boldsymbol{u}$ reliably in spite of the errors on his channel!

References

[1] A. D. Wyner, "The wiretap channel," *Bell Syst. Tech. J*, vol. 54, pp. 1355–1387, Oct. 1975.

[2] T. M. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 2–14, Jan. 1972.

[3] E. C. van der Meulen, "A survey of multiway channels in information theory," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 1–37, Jan. 1977.

[4] A. B. Carleial and M. E. Hellman, "A note on Wyner's wiretap channel," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 387–390, May 1977.

[5] S. K. Leung-Yan-Cheong, "Multiuser and wiretap channels including feedback," Ph.D. thesis, Dep. of Electrical Engineering, Stanford Univ., July 1976.

[6] I. Csiszar and J. Körner, "Broadcast channels with confidential messages," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 339–348, May 1978.

[7] S. K. Leung-Yan-Cheong, "On a special class of wiretap channels," *IEEE Trans. Information Theory*, vol. IT-23, pp. 625–626, Sept. 1977.

[8] S. K. Leung-Yan-Cheong and M. Hellman, "The Gaussian wiretap channel," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 451–456, July 1978.

[9] R. Kahn and M. Hellman, "On the wiretap channel with feedback," in preparation.

[10] E. Parzen, *Stochastic Processes*. San Francisco: Holden-Day, 1962.

[11] L. A. Shepp and D. Slepian, "An interesting elementary binary stochastic process," Bell Telephone Laboratories Tech. Mem. 20878–4, April 17, 1963.

[12] D. S. Ornstein, "An application of ergodic theory to probability theory," *Ann. Prob.*, vol. 1, pp. 43–65, Feb. 1973.

[13] M. E. Hellman, "Convolutional source encoding," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 651–656, Nov. 1975.